# Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis

Year: 2023 | Citations: 526 | Authors: Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu

---

## Abstract

Recent text-to-image generative models can generate high-fidelity images from text inputs, but the quality of these generated images cannot be accurately evaluated by existing evaluation metrics. To address this issue, we introduce Human Preference Dataset v2 (HPD v2), a large-scale dataset that captures human preferences on images from a wide range of sources. HPD v2 comprises 798,090 human preference choices on 433,760 pairs of images, making it the largest dataset of its kind. The text prompts and images are deliberately collected to eliminate potential bias, which is a common issue in previous datasets. By fine-tuning CLIP on HPD v2, we obtain Human Preference Score v2 (HPS v2), a scoring model that can more accurately predict human preferences on generated images. Our experiments demonstrate that HPS v2 generalizes better than previous metrics across various image distributions and is responsive to algorithmic improvements of text-to-image generative models, making it a preferable evaluation metric for these models. We also investigate the design of the evaluation prompts for text-to-image generative models, to make the evaluation stable, fair and easy-to-use. Finally, we establish a benchmark for text-to-image generative models using HPS v2, which includes a set of recent text-to-image models from the academic, community and industry. The code and dataset is available at https://github.com/tgxs002/HPSv2 .